Social or Individual Disagreement? Perspectivism in the Annotation of Sexist Jokes

Berta Chulvi[1,*], Lara Fontanella[2], Roberto Labadie-Tamayo[1] and Paolo Rosso[1]

[1] *Universitat Politècnica de València*
[2] *G. d'Annunzio University of Chieti-Pescara*

Abstract
The purpose of this paper is to show that the disagreement expressed in the data does not come from individual differences but from diverse and sometimes conflicting, social positions. Using a medium size dataset, 210 sexist jokes and 76 annotators, we test the hypothesis that, from a certain point (size of 12 in our data), adding more subjects to the annotation process does not increase the disagreement. We also measure the attitudes of subjects in sexism, introducing a new scale of *Hostile Neosexism*, and the consistent or inconsistent behaviour of annotators regarding their attitudes. We propose that perspectives are a combination of attitudes and behaviours, and we explore how they affect inter-rater agreement and which will be the number of annotators that we need to include all the perspectives in an annotation strategy.

1. Introduction

Artificial Intelligence (AI) applications often perpetuate and accentuate unfair biases that can originate from multiple sources, such as data sampling, labelling processes, training data, etc. This paper focuses on new strategies for reducing bias in the labelling process following the Learning from Disagreements paradigm (for a recent review, see [1]). This new approach in Natural Language Processing (NLP) tries to avoid the bias of considering a unique and correct vision of one phenomenon captured by a gold standard corpus, even when the problem addressed is the object of a strong social debate such as hate speech or sexist language. The research we present raises two fundamental questions, one of a theoretical nature - what is the nature of these disagreements that we need to consider? - and the other of a methodological nature: how to approach an annotation process that includes the different perspectives of a phenomenon considering the existence of limited resources for the labelling process?

Regarding the first theoretical question, in social psychology there is strong evidence that humans disagree even in seemingly objective tasks like estimating which line has the same length as a standard line [2, 3]. It has been studied in detail how these disagreements do not occur in a social vacuum due to individual differences in perception, but instead are the result of social influence strategies with implications for the individuals at the level of their social relations or their social identity (for a recent review of this literature, see [4]). In line with this tradition of research, the present study tries to demonstrate that the Learning from Disagreement paradigm needs to consider disagreement as a social phenomenon and not at the individual level. Individual attitudes towards various issues, such as equality, abortion, or immigration, are the expression of ideological and social conflicts in which individuals take part. Then, the general idea underlying this research is that when dealing with socially relevant problems, NLP tasks need to consider that different perspectives in the data respond to different social positions in the social realm. The hypothesis derived from this assumption is that from a certain point on, the inclusion of more individuals in an annotation process does not produce more disagreement [H1]. If the results verify this hypothesis, the following research question is how to estimate the optimal size of a group of annotators from which disagreement does not change significantly [RQ1].

To identify bias in the labelling process, recent research in NLP focuses on demographic, ideological, and attitudinal differences among individuals [5]. We propose that considering only attitudes and ideology is insufficient to approach the perspectivism paradigm correctly. A characteristic of human beings that we know from the beginning of social psychological research is that attitudes do not always predict behaviour [6] or do not directly predict behaviour [7]. People's inclination for consistency is widely acknowledged, and while they occasionally manage to maintain it, more often than not, they fall short of achieving it. Social psychology has developed a vast theoretical and empirical effort to understand consistency and inconsistency in human attitudes and behaviour [8, 9]. As labelling is a behaviour, a second assumption arising from our research is that different perspectives in annotation will be related not only to the expression of certain attitudes but also to the fact of acting consistently or inconsistently with the values these attitudes express.

2nd Workshop on Perspectivist Approaches to NLP

[*] Corresponding author.

berta.chulvi@upv.es (B. Chulvi)
0000-0003-1169-0978 (B. Chulvi); 0000-0002-5441-0035 (L. Fontanella); 0000-0003-4928-8706 (R. Labadie-Tamayo); 0000-0002-8922-1242 (P. Rosso)

Then a hypothesis derived from this assumption is that agreement in an annotation process will change considering individuals' attitudes related to the issue and the consistent or inconsistent annotators' behaviour in the annotation process [H2]. If the results verify this hypothesis, the research question is which size of the annotators' group ensures that our annotators' team reproduce the mix of perspectives that reflect well attitudes and the consistent or inconsistent behaviour with them, which gives the complete picture of a controversial debate [RQ2].

Using a relatively small corpus (210 sexist jokes) and a large group of 76 annotators, we test hypotheses 1 and 2 and try to answer the two research questions about which will be the optimal size of the group to include different perspectives [RQ1] and how to ensure our annotators reproduce a representative mix of perspectives [RQ2].

The rest of the article is organised as follows. Section 2 presents previous research related to the concepts that we use. In Section 3, we present our empirical research: data, task, and procedure. Details about the statistical analyses are given in Section 4. We present the results of our empirical evaluation in Section 5 and conclusions and limitations in Section 6.

2. Related work

2.1. The perspectivism sift and the labelling bias

In modern computational linguistics, the standardised annotation process of a corpus includes different techniques to classify a single piece of language in a given taxonomy. It implies training annotators, multiple classification subjects, measures of inter-annotator agreement, harmonisation, aggregation by the majority, and construction of a "gold standard" corpus representing the truth against which future predictions of NLP models will be compared. According to the tasks' taxonomy of Perez and Mugny [10], it means that the labelling process is being approached as an *aptitude task*, that is, a task with a correct answer (see Figure 1). This approach is hardly applicable when confronted with what different authors have referred to as a "highly subjective task" [11, 12]. We propose to denominate these tasks *opinion tasks*, following the taxonomy of [10], because their main characteristic is not their subjectivity but the fact that, looking at the way that society considers them, it seems that a correct answer does not exist (low relevance of error). Still, all the possible answers situate the person at the point of a continuum whose extremes are defined by a social confrontation (high social relevance). We view the sift paradigm, proposed in the *Perspective Data Manifesto*[1], as a more stringent approach to handling opinion tasks in NLP. The sift paradigm advocates for the publication of datasets in pre-aggregated form and the development of new measures for the evaluation of models that take into account all the perspectives linked to different backgrounds.

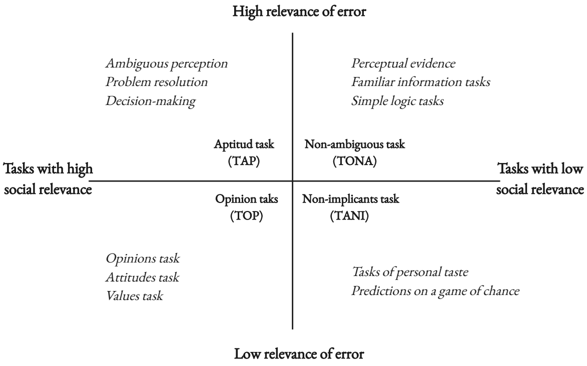

Figure 1: Taxonomy of task by Perez and Mugny (1993) and examples.[2]

The research adopting perspectivism in NLP grows year by year (for a recent review, see [1]) and one main concern is the labelling bias introduced by the cultural background of annotators [13, 14].

In recent research, Sap and colleagues [5] have shown strong associations between annotator identity and beliefs and their ratings of toxicity. Specifically, their results show that more conservative annotators and those who scored highly on a racist beliefs scale were less likely to rate anti-black language. Closer to our research questions is the work of Akhtar et al. [15, 16], which leverages different opinions emerging from groups of annotators with the goal of studying how polarised instances affect the performance of the classifiers. Considering binary classification tasks, they introduce a novel measure of the polarisation of opinions able to identify which instances in a dataset are more controversial. In a pilot study about xenophobia arguments in the context of Brexit, the annotation process was organised to contrast the annotation done by three people with an immigrant background (target group) in front of three people with a mainstream background as a control group. Using their polarisation index, the authors show how in several tweets, all the members of the target group (immigrants) marked the message as racist and hateful, while the members of the control group marked it as conveying no hate or racism. It is interesting to note that they only found a few tweets (1.13%) on which all the annotators agreed that they contained hateful messages. Implicitly, in this work the authors assume, similar to our perspective, that the nature of the disagreement is social and sustained by a social conflict, but they do not provide any empirical

[1] https://pdai.info/

[2] For the tasks classification, we have kept the original acronyms from the French version.

measure of annotators' attitudes. Their results suggest that consensus-based methods to create gold standard data are not necessarily the best choice when dealing with what they call highly subjective phenomena and we consider *opinion tasks*.

2.2. Attitudes and behaviour relation

In binary classification tasks, annotating a corpus is a behaviour more than the expression of an opinion. The annotators will use their attitudes and beliefs to decide, but it is hard to expect that attitudes predict perfectly this behaviour. Attitudes influence behaviour, as we have already seen in the work of [5], but the relation attitude-behaviour is not a pacific question in social-psychology literature (for a classical review, see [17]). For example, Donald Campbell [18], in the sixties, argued that people who hold negative attitudes toward minorities may be reluctant to express their attitudes through public behaviour because norms of tolerance and politeness were typically held in American society. Things have changed a lot regarding the open expression of hate towards minorities, that is why The New York Times published, in 2019, an editorial with the suggestive headline of "Racism Comes Out of the Closet"[3].

Not only does agreeing with social norms and situational constraints explain the inconsistencies between attitudes and behaviours, but there are also specific domains, such as humour, that significantly facilitate these kinds of inconsistencies. Often, some groups use humour to avoid moral judgement that penalises discrimination. Offensive people find support from a majority who consider that some messages are "only" jokes. When a society begins to overcome its prejudices towards certain social groups, we can observe that humour becomes a space in which these prejudiced attitudes are maintained. In fact, when we examine offensive jokes, we find they are mainly related to some social minorities [19]. These inconsistencies between attitudes and the behaviour of the annotators could also be a symptom of changes or resistances of subjects and capture the evolution of some opinion groups in controversial debates.

2.3. The Hostile Neosexism

Traditionally, sexism [20] has been viewed as the holding of discriminatory attitudes toward women, both manifest and subtle. This distinction in the tone of sexism was proposed by the ambivalent sexism theory [21, 22]. It was developed to account for a sort of evolution from a hostile component of sexism (overtly negative attitudes towards women) to a benevolent component (attitudes towards women that seem subjectively positive but are actually discriminatory). These two components differ in tone but are positively correlated and work together to perpetuate gender inequalities (for a recent review, see [23]). Also related to the evolution of sexism, is the concept of neosexism [24] or modern sexism [25]. Like modern racism, modern sexism is characterised by the denial of continued discrimination, antagonism toward women's demands, and lack of support for policies designed to improve women's position in society.

In a recent review on ambivalent sexism, Barreto and Doyle [23] point out future directions in the study of sexism due to the rapid developments in societal norms and attitudes towards sex, gender, and sexuality across many countries. Surprisingly, despite an important amount of research noting a rise in the number of men with a self-proclaimed anti-feminist agenda [26, 27, 28], these authors do not consider as future work to investigate the link between hostile sexism and anti-feminist attitudes. To go deeper into the interaction between hostile sexism and anti-feminist attitudes seems relevant because a new kind of strong hostility towards women uses anti-feminist frames, but also supports certain feminist policies, such as equality [29]. This new latent attitude, that we denominate *Hostile neosexism*, is difficult to capture with old attitudes scales towards feminism, such the one developed by Smith in the seventies [30], because most of the items of this instrument fit with the feminist values that this new *Hostile neosexism* seems to support. Also, it seems to get out from the scope of the whole ambivalent sexism inventory [21] that does not pay specific attention to feminism itself. Regarding the modern sexism scale [25] or the Neosexism scale [24], we argue that *Hostile neosexism* presents a high degree of hostility against women that the previous scales do not capture[4]. The core of this *Hostile neosexism* attitude is the claim that societal changes driven by the feminist movement are inherently unfair and put men as a group in a disadvantageous position. Despite, the hostile sexism subscale [21] was primarily driven by the idea that men's dominance over women is both appropriate and desirable, some items of this subscale connect well with the idea that nowadays there is no reason for feminist demand and that the feminist movement overreacts (see items 3, 4 and 5 in Section 3.2.1).

3. Study Design

3.1. Data

To carry out our study, we relied on a manually selected set of 210 jokes, conveying prejudice against women, from the corpus proposed in the shared task: HUrtful HUmour (HUHU): Detection of Humour Spreading Preju-

[3] https://www.nytimes.com/2019/07/15/opinion/trump-twitter-racist.html

[4] Authors are currently conducting research to test the need for this new instrument and validate a longer version of the scale

dice in Twitter at IberLEF 2023 [31]. This dataset offers a gold standard corpus of tweets in Spanish containing prejudice against four minorities: women, the LGBTIQ community, immigrants and racially discriminated people, and overweight people. During the annotation process of the HUHU dataset each instance was assessed for the presence of humour and prejudice by 3 annotators. The criterion used for annotation was based on the relative majority agreement of the annotators, with a threshold of 2 out of 3. For the present study, we select jokes that convey different kinds of prejudice against women. We have classified the 210 jokes into 5 categories with the aim of describing the content of the dataset providing some examples:

1. **Present women as dummies, only concerned about their bodies or about money** (40% of the dataset), e.g.: "If Socrates had been a woman he would have said: "I just know that I don't know what to wear".
2. **Feature women as possessive, complicated and dominant** (22.5%), e.g: "Women get angry for 5 reasons: 1) For everything 2) For nothing 3) Because they do 4) Because they don't 5) Just in case".
3. **Say that they are gossips and enemies among themselves** (2.5%), e.g: "If women governed, there would be no World War III, only little groups of countries badmouthing and smiling at each other"
4. **Introduce them as malicious, sluts or justifies violence** (12%), e.g: "Women are like bags of ice, with a few punches they loosen up"
5. **Anti-feminist jokes** (23%), e.g: "I have just been informed that Spanish troops on the war front are being brutally offended by macho and patriarchal attitudes on the part of the Russian army. It is a disgrace that this is still happening in the 21st century."

3.2. Participants and procedure

A total of 76 students of psychology (76.3% women and 23.7% men) took part in the experiments as an activity of a practical workshop in the first year of the degree. The activity was done in silence without other any distractions and took two hours time. Students were assigned a secret number to keep anonymity and access an Excel document to label the jokes. Annotation of task 1 consisted in reading the 210 jokes and classifying them as sexist (containing a prejudice against women) or not. The annotators had to say also whether the text contains humour or not (task 2) and which was the offensiveness of prejudice (task 3) on an ordinal scale (0=not at all, 1=slightly; 2=somewhat, 3=very much). After completing the annotation task, using the secret number, students responded to a questionnaire containing the *Hostile neosexism* scale and a question about their ideology.

3.2.1. Annotators attitudes and ideology

To measure annotators' attitudes in *Hostile Neosexism*, we created a short scale that we denominate *Brief Hostile Neosexism Scale.* It is composed of six items: three of them (4 to 6) are part of the *Hostile Sexism subscale* of the *Ambivalent Sexism Scale* from Glick and Fiske [32] and the other three (1 to 3) are new items that we created *ah hoc* to measure anti-feminist attitude:

1. Some of the demands of the feminist movement seem to me to be a bit exaggerated.
2. I sometimes feel that our society pays too much attention to the rights of certain minorities.
3. In the name of equality, many women try to gain certain privileges.
4. Many women interpret innocent comments and actions as sexist.
5. Women are easily offended.
6. Women exaggerate the problems they suffer because they are women.

As political conservatism has been found to explain more variance in *ambivalent sexism* than gender [33, 34], we asked annotators to indicate their ideological position. The question was the following: If you had to define your political orientation, where would you place yourself on this scale? The answer must be expressed on a 7-point Likert-type scale where 1 was "left" and 7 was "right" [5]. The voluntary participation and the anonymity of data were guaranteed following the European Code of Conduct for Research Integrity[6].

4. Statistical analyses

As discussed in the Introduction, our research aims to evaluate the influence of attitudes on the annotation process and the relation between attitudes and behaviour. To derive annotators' latent attitude and behaviour, we exploit an Item Factor Analytic approach, which constitutes an extension of classical linear factor analysis and is particularly suitable for addressing categorical variables. Specifically, within the framework of Item Response Theory (IRT) [35], we adopt the two-parameter normal ogive (2PNO) formulation [36]

$$Pr(X_{ik}=c|\theta_i,\gamma_k,\lambda_k) = \Phi\left(\lambda_k\theta_i-\gamma_{k,c}\right) - \Phi\left(\lambda_k\theta_i-\gamma_{k,c+1}\right) \text{‘} \quad (1)$$

[5] Data are public in https://github.com/Bertachulvi/ECAI2023

[6] https://allea.org/code-of-conduct/

where $\Phi(\cdot)$ is the normal cumulative function. Here the probability of observing a given category $c = 1, \dots, C$, for unit $i = 1, \dots, N$ and item $k = 1, \dots, K$, is modelled in terms of the latent trait θ_i, the factor loading λ_k and a vector of ordered threshold $\boldsymbol{\gamma}_k$. To estimate the model parameters, we embrace a fully Bayesian approach that incorporates the handling of missing values [37].

We are also interested in measuring inter-rater agreement in the task of annotating sexism. As expected, because our data come from the HUHU dataset, we have observed that in the binary annotation scheme, most of the texts are categorised as jokes conveying prejudice against women, with 81% of the annotations falling into this category. This skewed distribution of data leads to a low level of agreement among different raters when using traditional inter-rater agreement measures such as Fleiss' κ or Kripendorf's α. This discrepancy arises from the paradoxical situation where the observed agreement appears to be very high, while the chance-corrected agreement is actually low [38]. To address this issue, we employ Gwet's AC$_1$ measure of inter-rater agreement [39], which utilises a probabilistic model of agreement [40]. This approach estimates the difficulty levels of the items within the corpus through probabilistic inference and then estimates the probability of chance agreement separately for easy and hard items. This probabilistic modelling approach helps mitigate the impact of the skewed data distribution on the agreement assessment process.

5. Results

5.1. Do more annotators produce more disagreement?

To test hypothesis 1 which considers disagreement as a social phenomenon and not at the individual level, we need to investigate the influence of the number of annotators on inter-rater agreement. For doing so, we randomly selected samples without replacement from the population of 76 annotators, with sample sizes n ranging from 3 to 45. To ensure statistical robustness, 10,000 iterations were performed for each sample size. The results of this analysis are presented in Figure 2. In particular, Figure 2(a) depicts the mean and 95% confidence interval for each sample size. To determine the optimal annotator sample size that leads to stabilisation in the variability of Gwet's AC$_1$ coefficient, the knee-point method was employed [41]. This method is commonly used to identify the point at which a graph exhibits a significant change in slope. In this study, the knee-point method was applied to the amplitude of the confidence intervals (see Figure 2(b)). Through the application of the knee point method, an annotator sample size of $n = 12$ was determined to be the point of stabilisation for AC$_1$ variability, indicating that further increases in the number of annotators do not yield significant modification in agreement [RQ1].

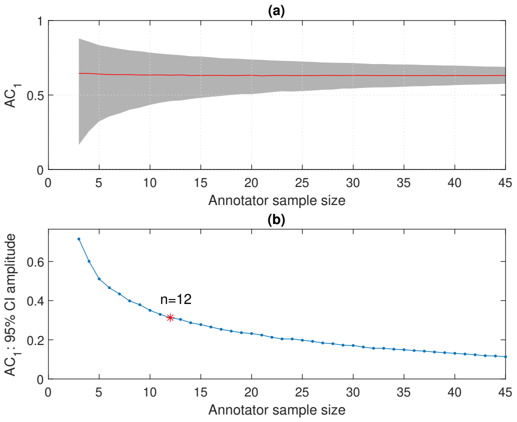

Figure 2: Simulation results: (a) Mean and 95% confidence interval of Gwet's AC$_1$ coefficient; (b) Amplitude of the 95% confidence interval of Gwet's AC$_1$ coefficient and knee-point.

5.2. How do attitudes affect the agreement among annotators?

A Bayesian exploratory IRT analysis was employed, following the approach described in [42], in order to evaluate the construct validity of the scale outlined in Section 3.2.1. The results of the analysis indicated that the scale exhibits unidimensionality, supporting its validity as a measurement tool for the intended construct. Therefore, a unidimensional 2PNO model (Equation 1) was exploited to estimate the *Hostile neosexism* attitude of the annotators, taking into account the influence of their gender and ideology as relevant features. The estimated values for the model parameters can be found in Table 1. The factor loadings indicate the weight of the corresponding items in the derivation of the latent trait scores, while the location values give insights on the level of consolidation of the corresponding *Hostile neosexism* attitude: lower values correspond to a belief that gains more support in our sample [43]. As for the regression parameter estimates, the only covariate that seems to significantly impact the *Hostile neosexism* attitude is endorsing right ideology.

To assess the influence of the *Hostile neosexism* attitude on the level of agreement, we contrast the inter-rater agreement among the $n = 12$ annotators in three subgroups: a homogeneous group with the lowest scores on the *Hostile neosexism* attitude, a homogeneous group with the highest scores, and a mixed group with six annotators positioned at the lower end of the *Hostile neosexism*

Table 1
Hostile Neosexist scale: parameter estimates

	posterior mean	95% credible interval
	Factor loadings	
Item6	1.674	(1.054, 2.671)
Item1	1.137	(0.745, 1.572)
Item3	1.059	(0.706, 1.433)
Item4	1.408	(0.950, 1.996)
Item5	1.271	(0.826, 1.821)
Item2	0.717	(0.404, 1.034)
	Location value$^{(a)}$	
Item1	-0.971	(-1.320, -0.619)
Item3	-0.640	(-0.948, -0.333)
Item4	-0.536	(-0.886, -0.207)
Item2	0.444	(0.157, 0.733)
Item5	0.682	(0.390, 0.989)
Item6	1.131	(0.761, 1.541)
	Regression coefficients	
intercept	-0.628	(-1.256, 0.000)
male	0.351	(-0.244, 0.946)
left$^{(b)}$	-0.500	(-1.217, 0.215)
moderate left$^{(b)}$	0.000	(-0.753 , 0.733)
right$^{(b)}$	0.744	(0.004, 1.513)

(a) average of the threshold values for a given item
(b) baseline: centre

and six annotators positioned at the higher end. The observed and expected agreements and the Gwet's AG$_1$ coefficients for all the 76 annotators and for the 3 subgroups are displayed in Table 2. The results demonstrate a clear distinction in the level of agreement among the annotators with lower *Hostile neosexism* attitude compared to the other groups. On the other hand, the agreement within the mixed group is similar to that observed in the overall population of annotators, indicating a comparable level of consensus among individuals with varying levels of *Hostile neosexism* attitude.

Table 2
Inter-rater agreement comparison for samples with different Hostile Neosexism attitudes

	n	observed agreement	expected agreement	Gwet's AC1$_1$
All annotators	*76*	*0.741*	*0.298*	*0.631*
Lowest Hostile Neosexism	12	0.828	0.209	0.782
Highest Hostile Neosexism	12	0.722	0.306	0.599
Mixed Hostile Neosexism	12	0.751	0.273	0.658

We develop a second sub-sampling strategy to test the influence of attitudes on the level of agreement. A simulation was conducted with a sample size of $n = 12$, and the sample units were randomly selected from subpopulations characterised by scores on the latent trait below the first quartile (*Low Hostile Neosexism*), above the third quartile (*High Hostile Neosexism*), and evenly distributed between the two sub-populations (*Mixed Hostile Neosexism*). From each group, we selected 10,000 samples without replacement. The findings (see Figure 3) provide further evidence of the influence of attitude on the level of agreement in the annotation process.

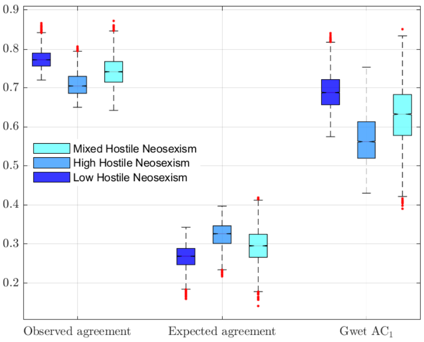

Figure 3: Observed and Expected inter-rater agreements and Gwet's AC$_1$ for samples of $n = 12$ annotators with low, high and mixed scores on the Hostile Neosexism attitude.

Following the two strategies, we find that the level of agreement decreases among the *Mixed Hostile Neosexism* group but also among *High Hostile Neosexism*. The decline in agreement among mixed groups is understandable but would not be expected among homogeneous groups high in *Hostile Neosexism*. Then we address the inconsistency between attitude and behaviour discussed in Section 2.2.

5.3. Are attitudes consistent with the annotators' behaviour?

An alternative approach based on IRT models, as proposed in [44], can be employed to gain insights into consistency in annotators' behaviour across the 210 tweets, specifically regarding their ability to recognise instances of sexism in the jokes. This alternative formulation of the IRT model deviates from the traditional approach by treating the annotators as items, allowing the threshold parameter in the binary annotation task to be interpreted in terms of the level of difficulty in recognising the presence of classical sexist content in jokes[7]. We denominate this variable *Sexism Recognition Shortcoming* because all text comes from a dataset that expresses sexism, but we do not interpret these recognition problems as a lack of skill, but rather, as the expression of an opinion. As

[7] We use the classical adjective here because a 77% of jokes refer to traditional misogynistic stereotypes that present women as dumb, body-centred, gossipy, incomprehensible for men or malicious

the pragmatic of communication emphasises, every behaviour is a communication act, even the silence [45].

As depicted in Figure 4, there is evidence of a positive correlation between the *Hostile Neosexism* attitude of the annotators and their *Sexism Recognition Shortcoming* behaviour, reinforcing the idea that attitude and behaviour are connected. However, the intriguing result is that the strength of this association is relatively modest, as indicated by the Pearson's correlation coefficient (ρ = 0.234). This suggests that the impact of attitude on the behaviour of identifying the presence of sexist content is somewhat limited and we need to introduce a more complex view to identify the different perspectives.

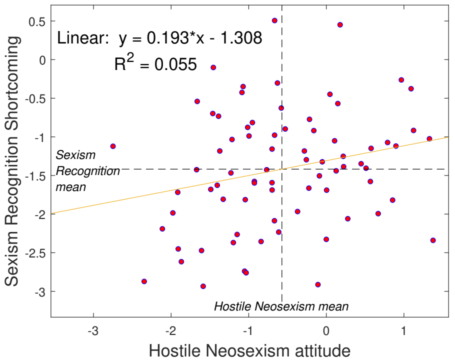

Figure 4: Relation between Hostile Neosexism attitude of annotators and Sexism Recognition Shortcoming behaviour.

To further explore the relationship between attitude and behaviour, we classified the annotators into four groups based on their positioning relative to the means of the two identified variables: *Hostile Neosexism* attitude and *Sexism Recognition Shortcoming* (see Table 3). As we can see, the most numerous are the consistent groups: low-low (34%) or high-high (27%). However, the number of individuals exhibiting annotation behaviour inconsistent with expressed attitudes (22.4% and 15.8%) is not negligible.

Table 3
Groups' composition according to Hostile Neosexism attitude and Sexism Recognition Shortcoming

	Sexism Recognition Shortcoming	
Hostile Neosexism	Low	High
Low	26	17
	34.2%	22.4%
High	12	21
	15.8%	27.6%

This grouping allows for a more nuanced examination of how different positions on the attitude and behaviour dimensions may be related to some annotators' characteristics. Table 4 provides the percentage composition of the identified groups in terms of gender and ideology. The chi-square test of independence leads to conclude that there is a significant association between those characteristics and the group identified along the sexist latent traits (gender: p-value 0.0018; ideology: p-value 0.0014). As we can see, the expected result on the impact of gender and ideology showed in Table 4 are especially manifest in consistent groups. The left is the majority in Low-Low group, and the right in the High-High group. The novelty is that we can mostly link the inconsistencies with the moderate left. This group finds different partners in the inconsistency behaviour: the left in the *low Hostile Neosexism-high Sexism Recognition Shortcoming* (Low-High) group and the right in the *high Hostile Neosexism-low Sexism Recognition Shortcoming* (High-Low) group.

Table 4
Gender and Ideology in the different groups of annotators

	Total	Low	Low	High	High
		-	-	-	-
		Low	High	Low	High
Male	23.7%	11.5%	11.8%	25.0%	47.6%
Female	76.3%	88.5%	88.2%	75.0%	52.4%
Left	36.8%	53.9%	41.2%	25.0%	19.0%
Moderate left	22.4%	15.4%	41.2%	33.3%	9.5%
Centre	19.7%	19.2%	17.6%	8.4%	28.6%
Right	21.1%	11.5%	0.0%	33.3%	42.9%

With the inclusion of two supplementary annotation tasks as outlined in Section 3.2, we can assess whether the inconsistencies among annotators are related to the perception of humour in tweets or to their judgement of the level of offensiveness associated with each text. To this end, we used a procedure similar to the one described in Section 5.2 in order to derive annotators' scores on the latent dimensions of *Humour recognition* and *Degree of offensiveness.* Figure 5 shows the distribution of the estimated scores for the recognition of humorous content and for the evaluation of the degree of offensiveness across the four annotators' groups.

In Figure 5, we appreciate that the inconsistency between attitudes and behaviour in the case of individuals with *Low Hostile Neosexism* attitude but *High Sexism Recognition Shortcoming* relies on a higher recognition of the text as humorous. This inconsistency supports the implicit and extended assumption that humour does not hurt. This group is also the one that rates tweets as less offensive. In this group, the left and the moderate left represents the 82.4% of the total. Humour recognition also plays a role in the other inconsistent group, the individu-

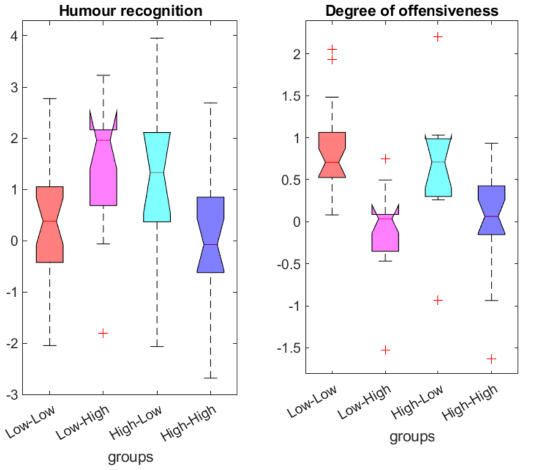

Figure 5: Humour recognition and Assessment of the offensiveness level in the different groups of annotators.

als with *High Hostile Neosexism* attitude but *Low Sexism Recognition Shortcoming* where moderate left and right sum to 66.6%. We believe that this group, with its inconsistency, is expressing that annotators embrace *Hostile Neosexism* which targets the feminist movement as overacting but recognises well the classical sexism expressed in 77% of jokes. For this interpretation, it is important to take into account that our data mostly fits with categories that express classical prejudices and stereotypes against women (see Section 3.1). The position of the two consistent groups (Low-Low and High-High) seems coherent: for different reasons, some because jokes contain prejudice (Low-Low), others because maybe they think jokes describe reality well (High-High), both find the tweets less humorous, but they differ in the degree of offensiveness. As expected, for the High-High group tweets are less offensive than for the Low-Low group. These results lead us to affirm that perspectives are expressed through a combination of attitudes and behaviours.

5.4. Agreement and perspectives

In this section, we explore whether the agreement changes considering individual's attitudes and consistent or inconsistent behaviour [H2]. As we see in Table 5, individuals with similar attitude, *Low Hostile Neoexism*, will exhibit very different inter-rater agreement (0.83 > 0.37) if we consider the consistency between attitudes and behaviour. The same occurs with the opposite attitude: *High Hostile Neosexist* people exhibit very different inter-rater agreement (0.82 > 0.49) if we consider the consistency between attitudes and behaviour.

We can not conclude that an inconsistent behaviour reduces the agreement because, in the *Low Hostile Neosexism* group, high agreement occurs in the consistent subgroup, while in the *High Hostile Neosexism* group, it occurs in the inconsistent subgroup. As we argue in Section 5.3, individuals communicate their opinions not only through attitude expression but also through behaviour, as the pragmatics of communication assesses [45]. In this regard, we interpret high inter-rater agreement as the identification of a clear social position and low inter-rater agreement as the existence of a changing social position. By changing social position we mean a process in which individuals did not find a clear indication in the social realm about which will be the action that must be expected from them in the given context. Then, the interpretation of the different perspectives must focus on identifying which kind of consensus or conflict causes the respective high or low agreement. We do not think that different perspectives must be matched with different groups with a strong agreement because not polarised groups on a particular issue could exhibit a low level of agreement (according to what [15] propose). This group might also express a different perspective as a way to approach a controversial issue even if there is not a polarised position, because this lack of polarisation is what defines the group. Moreover, we need to consider controversial issues dynamically, and then it is reasonable to think that new perspectives, or changing ones, will register low levels of agreement because they reflect a social position that is being formed or one that is in crisis. Our interpretation of the different perspectives that we find in our data, taking into account the nature of the task of labelling a corpus that entirely contains sexist jokes, is the following:

Table 5
Inter-rater agreement across the groups according to Hostile Neosexism attitudes and Sexism Recognition Shortcoming

Hostile Neosexism	Sexism Recognition Shortcoming	observed agreement	expected agreement	Gwet's AC_1
Low	Low	0.865	0.167	0.838
Low	High	0.648	0.435	0.377
High	Low	0.857	0.160	0.829
High	High	0.676	0.361	0.493

1. **Low-Low group**: People that highly support the modern feminist movement (*Low Hostile Neosexism*) and that do not find funny (*Low Sexism Recognition Shortcomings*) classical sexist jokes. It is a clear social position in sociological terms, then we find a high agreement (Gwet's AC_1=0.838).
2. **Low-High group**: People that support the modern feminist movement (*Low Hostile Neosexism*) but still find funny (*High Sexism Recognition Shortcomings*) classical sexist jokes. It is a changing

social position in sociological terms because the mainstream message is that this humour is not funny, then we find a lower agreement (Gwet's AC_1=0.37).

3. **High-Low group**: People that do not support the modern feminist movement (they think that some feminist overreacts) but give support to the old feminist movement (the one that emphasises equality) and is able to recognise offensiveness in the sexist jokes. This is a clear social position because fits with the 20th century feminism, and then we find a high level of agreement (Gwet's AC_1=0.829).
4. **High-High group**: People that represent new phenomenon that we have labelled as *Hostile Neo-sexism.* They manifest a strong hostility to the modern feminist movement that could lead to a not recognition of the classical sexism jokes, that is, it can endanger the achievements of the equality movement during the 20th century. This a new social position and then we find a low level of agreement (Gwet's AC_1=0.49).

Aside from the aforementioned understanding of the various views, we believe that multiple perspectives should be be present in an ideal team of annotators. The next study research question is about determining the ideal size of the group to include all of them based on our data.

5.5. Size of the group and perspectives

Assuming the composition of the annotators' population detailed in Table 3, our objective in this section is to investigate the sample size required to ensure the inclusion of all diverse perspectives within an annotator team [RQ2]. To achieve this, we randomly selected, with replacement, 100 samples from the original population for each sample size in the range 2-45. The representativeness of each sample with respect to the composition of the original population was assessed using the Frobenius distance between the original and the sample composition. The knee-point method was employed to identify the optimal sample size, meaning the sample size that guarantees a minimal distance between the sample and the population composition in terms of the proportion of annotators belonging to the four identified groups. To ensure the robustness of our findings, we repeated the simulation procedure 1000 times, resulting in an empirical distribution of the optimal sample size across the repetitions (see Table 6). From the results, we can conclude that for our study a sample size ranging from 10 to 12 will most likely guarantee a fair representation of the different perspectives in the annotators' team.

Table 6
Empirical distribution of the optimal sample size

Optimal sample size	9	10	11	12	13
Number of repetitions	11	229	605	138	17
% over repetitions	1.1%	22.9%	60.5%	13.8%	1.7%

6. Conclusion and limitations

In this paper, we presented a methodology that approaches several common problems that arise when we intend to translate the perspectivism paradigm to a coherent annotation strategy. We tested H1, and our results in Section 5.1 suggest that the nature of the disagreement in the annotation is social and not individual because, from a certain point, it does not increase by adding more individuals. We apply a social psychology-grounded taxonomy for classifying tasks that could be helpful for dealing with what, in NLP research, is referred to as a subjective task. We also verify that different perspectives arise not only from attitudes but also from inconsistent or consistent behaviour of the annotators with these attitudes. We find this important because it shows that we can not assume that we will include all perspectives in a dataset only relying on attitude or biographical differences. We also argue that these inconsistencies are valuable information about how controversial issues evolve in social debate. We propose that perspectives are a combination of attitudes and behaviour. We evaluate which will be the size of the group to include all the perspectives detected in our data.

Several limitations of this work must be considered. First, the annotator team is composed of psychology students, but even within this homogeneous group, we have seen that different perspectives arise. Also, we choose to work with a dataset containing only sexist jokes, because we try to avoid the diversity coming from the data and to concentrate on annotators' perspectives, but a deep analysis of the text will give us more insights and a more complex view. The more challenging future work is to translate the knowledge obtained in this research into a feasible methodology to include all perspectives in an annotation plan that might need to proceed in three steps at the time of creating the corpus: (i) a first exploratory step that identifies perspectives and how these perspectives are reflected in the data, (ii) a second step to ensure the representativeness of the data in terms of perspectivism and (iii) a final step that control if, at the end of the annotation procedure, the data reflect all the perspectives.

Acknowledgments

Berta Chulvi and Paolo Rosso are supported by FairTransNLP-Stereotypes PID2021−124361OB-C31 funded by MCIN/AEI/10.13039/501100011033 and by ERDF, EU A way of making Europe. The work of Roberto Labadie was supported by valgrAI - Valencian Graduate School and Research Network of Artificial Intelligence and the Generalitat Valenciana. Lara Fontanella is supported by the ICOMIC (Identifying and Counteracting Online Misogyny in Cyberspace) Project funded by EU Next Generation, MUR-Fondo Promozione e Sviluppo-DM 737/2021

References

[1] A. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, Learning from Disagreement: A Survey, Journal of Artificial Intelligence Research 72 (2021) 1385–1470.

[2] S. E. Asch, Studies of independence and conformity: I. A minority of one against a unanimous majority, Psychological Monographs: General and Applied 70 (1956) 1–70. doi:10.1007/s11135-022-01494-7.

[3] J. D. Campbell, P. J. Fairey, Informational and normative routes to conformity: The effect of faction size as a function of norm extremity and attention to the stimulus, Journal of Personality and Social Psychology 57 (1989) 457–468. doi:https://doi.org/10.1037/0022-3514.57.3.457.

[4] R. Spears, Social Influence and Group Identity, Annual Review of Psychology 72 (2021) 367–390. doi:10.1146/annurev-psych-070620-111818.

[5] M. Sap, S. Swayamdipta, L. Vianna, X. Zhou, Y. Choi, N. A. Smith, Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 5884–5906. URL: https://aclanthology.org/2022.naacl-main.431. doi:10.18653/v1/2022.naacl-main.431.

[6] R. T. LaPiere, Attitudes vs. actions, Social forces 13 (1934) 230–237.

[7] I. Ajzen, M. Fishbein, Attitude-behavior relations: A theoretical analysis and review of empirical research, Psychological bulletin 84 (1977) 888.

[8] A. W. Kruglanski, K. Jasko, M. Milyavsky, M. Chernikova, D. Webber, A. Pierro, D. Di Santo, Cognitive consistency theory in social psychology: A paradigm reconsidered, Psychological Inquiry 29 (2018) 45–59.

[9] J. Cooper, Cognitive Dissonance: Where We've Been and Where We're Going, International Review of Social Psychology (2019). doi:10.5334/irsp.277.

[10] J. A. Pérez, G. Mugny, Influences sociales : la théorie de l'élaboration du conflit, 1993.

[11] V. Basile, It's the End of the Gold Standard as we Know it. On the Impact of Pre-aggregation on the Evaluation of Highly Subjective Tasks, in: DP@AI*IA, 2020.

[12] V. Basile, T. Caselli, A. Balahur, L. Ku, Editorial: Bias, subjectivity and perspectives in natural language processing, Frontiers in Artificial Intelligence 5 (2022). doi:10.3389/frai.2022.926435.

[13] M. Sap, D. Card, S. Gabriel, Y. Choi, N. A. Smith, The Risk of Racial Bias in Hate Speech Detection, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 1668–1678. doi:10.18653/v1/P19-1163.

[14] Z. Waseem, Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter, in: Proceedings of the First Workshop on NLP and Computational Social Science, Association for Computational Linguistics, Austin, Texas, 2016, pp. 138–142. doi:10.18653/v1/W16-5618.

[15] S. Akhtar, V. Basile, V. Patti, A New Measure of Polarization in the Annotation of Hate Speech, in: M. Alviano, G. Greco, F. Scarcello (Eds.), AI*IA 2019 – Advances in Artificial Intelligence, Springer International Publishing, Cham, 2019, pp. 588–603.

[16] S. Akhtar, V. Basile, V. Patti, Whose Opinions Matter? Perspective-aware Models to Identify Opinions of Hate Speech Victims in Abusive Language Detection, 2021. URL: arXiv:2106.15896v1[cs.CL]30Jun2021.

[17] A. H. Eagly, S. Chaiken, The psychology of attitudes, Harcourt brace Jovanovich college publishers, 1993, pp. 155–218.

[18] D. T. Campbell, Social attitudes and other acquired behavioral dispositions, in: S. Koch (Ed.), Psychology: A study of a science. Study II. Empirical substructure and relations with other sciences. Vol. 6. Investigations of man as socius: Their place in psychology and the social sciences, McGraw-Hill, 1963.

[19] L. I. Merlo, B. Chulvi, R. Ortega-Bueno, P. Rosso, When humour hurts: linguistic features to foster explainability, Procesamiento del Lenguaje Natural 70 (2023) 85–98.

[20] J. K. Swim, L. L. Hyers, Sexism, in: T. D. Nelson (Ed.), Handbook of prejudice, stereotyping, and discrimination, Psychology Press, 2009, p. 407–430.

[21] P. Glick, S. T. Fiske, The ambivalent sexism inventory: Differentiating hostile and benevolent sexism, Journal of personality and social psychology 70 (1996) 491.

[22] P. Glick, S. T. Fiske, Ambivalent sexism revisited, Psychology of women quarterly 35 (2011) 530–535.

[23] M. Barreto, D. Doyle, Benevolent and hostile sexism in a shifting global context, Nat Rev Psychol 2 (2023) 98–111.

[24] F. Tougas, R. Brown, A. M. Beaton, S. Joly, Neosexism scale, 1995.

[25] J. K. Swim, K. J. Aikin, W. S. Hall, B. A. Hunter, Sexism and racism: Old-fashioned and modern prejudices, Journal of Personality and Social Psychology 68 (1995) 199–214.

[26] M. Blais, F. Dupuis-Déri, Masculinism and the antifeminist countermovement, Social Movement Studies 11 (2012) 21–39. doi:10.1080/14742837.2012.640532.

[27] M. Blais, The impact of masculinist counter-framing on the work of meaning-making of violence against women, Interface 13 (2021) 353–382.

[28] K. M. O'Donnell, Incel mass murderers: Masculinity, narrative, and identity, Ohio Communication Journal 59 (2021) 64–76.

[29] G. Off, Complexities and Nuances in Radical Right Voters' (Anti)Feminism, Social Politics: International Studies in Gender, State & Society (2023). doi:10.1093/sp/jxad010.

[30] E. Smith, M. Ferree, F. Miller, A scale of attitudes toward feminism, Representative Research in Social Psychology 6 (1975) 51–56.

[31] R. Labadie-Tamayo, B. Chulvi, P. Rosso, Everybody Hurts, Sometimes. Overview of HUrtful HUmour at IberLEF 2023: Detection of Humour Spreading Prejudice in Twitter, in: Procesamiento del Lenguaje Natural (SEPLN), volume 71, 2023.

[32] P. Glick, S. T. Fiske, Ambivalent sexism inventory: Differentiating hostile and benevolent sexism, Journal of Personality and Social Psychology 70 (1996) 491–512.

[33] R. de Geus, E. Ralph-Morrow, R. Shorrocks, Understanding ambivalent sexism and its relationship with electoral choice in britain, British Journal of Political Science 52 (2022) 1564–1583.

[34] K. Hellmer, J. T. Stenson, K. M. Jylhä, What's (not) underpinning ambivalent sexism?: Revisiting the roles of ideology, religiosity, personality, demographics, and men's facial hair in explaining hostile and benevolent sexism, Personality and Individual Differences 122 (2018) 29–37.

[35] R. J. de Ayala, The Theory and Practice of Item Response Theory, The Guilford Press, New York, 2009.

[36] F. Samejima, Estimation of a latent ability using a response pattern of graded scores, Psychometric Society, Richmond, VA, 1969.

[37] L. Fontanella, P. Villano, M. Di Donato, Attitudes towards Roma people and migrants: a comparison through a Bayesian multidimensional IRT model, Quality and Quantity 50 (2016) 471–490.

[38] B. D. Eugenio, M. Glass, The Kappa Statistic: A Second Look, Computational Linguistics 30 (2004) 95–101. doi:10.1162/089120104773633402.

[39] K. L. Gwet, Computing inter-rater reliability and its variance in the presence of high agreement, British Journal of Mathematical and Statistical Psychology 61 (2008) 29–48. doi:10.1348/000711006X126600.

[40] S. Paun, R. Artstein, M. Poesio, Statistical Methods for Annotation Analysis, Springer Cham, Switzerland, 2022.

[41] D. Kaplan, Knee Point, https://www.mathworks.com/matlabcentral/fileexchange/35094-knee-points, 2023. [Online; accessed June 7, 2023].

[42] L. Fontanella, S. Fontanella, P. Valentini, N. Trendafilov, Simple Structure Detection Through Bayesian Exploratory Multidimensional IRT Models, Multivariate Behavioral Research 54 (2019) 100–112. doi:10.1080/00273171.2018.1496317.

[43] P. Villano, L. Fontanella, S. Fontanella, M. Di Donato, Stereotyping Roma people in Italy: IRT models for ambivalent prejudice measurement, International Journal of Intercultural Relations 57 (2017) 30–41. doi:https://doi.org/10.1016/j.ijintrel.2017.01.003.

[44] A. Tontodimamma, S. Fontanella, L.and Anzani, V. Basile, An Italian lexical resource for incivility detection in online discourses, Quality and Quantity (2022). doi:10.1007/s11135-022-01494-7.

[45] P. Watzlawick, J. B. Bavelas, D. D. Jackson, Pragmatics of human communication, WW Norton, 2011.